Audible Data Cleaning and Dashboard

1. Introduction:

The goal of this exercise was to emphasize the significance of data cleaning and analysis. To achieve this, the data cleaning process was conducted using Excel, and the subsequent dashboard analysis was carried out in Tableau.

2. Data Overview:

This section provides a detailed overview of the dataset used for analysis, including the attributes of audio books available on Audible. Each entry in the dataset represents an audio book, characterized by the following fields:

- **Name**: The title of the book.
- **Author**: The name of the author(s) of the book.
- **Narrator**: The name(s) of the narrator(s) who performed the audio book.
- **Time**: The total length of the audio book.
- **ReleaseDate**: The date on which the audio book was released on Audible.
- **Language**: The language in which the audio book is narrated.
- **Stars**: The average rating of the audio book, on a scale of 1 to 5 stars.
- **Price**: The price of the audio book in the relevant currency.

3. Data Cleaning Steps

3.1. Removing Duplicates

To ensure the uniqueness of each audiobook entry in our dataset, we employed Excel's "Remove Duplicates" feature found under the "Data" tab, applying it across all the fields. This process effectively identified and eliminated any duplicate records, leaving us with a dataset comprising only distinct audiobook entries.

3.2. Standardizing Formats

To streamline our dataset, we started by cleaning up the author and narrator columns using Excel's Find and Replace (CTRL + H) to remove "Writtenby:" and "Narratedby:". Next, we used Flash Fill to split full names into first and last names in the Author column, but encountered issues with entries having multiple narrators, leading to some data inconsistencies. To address this, we applied Power Query's technique for splitting columns based on capitalization (Microsoft documentation) and then combined the names again with the formula *“=CONCATENATE(A1, " ", B1)”*.

For the audiobook lengths, we unified the format by replacing "mins" with "min" and extracted the time data with these formulas:

- To get hours: *“=IF(ISNUMBER(SEARCH("hrs", D2)), VALUE(LEFT(D2, SEARCH(" ", D2)-1)), 0)”*
- To find minutes when "and" is present: *“=VALUE(MID(D2,SEARCH("and",D2)+4,SEARCH("min",D2)-SEARCH("and",D2)-5))”*
- If "and" is absent but "min" is found: *“=IF(ISNUMBER(SEARCH("min",D2)),VALUE(LEFT(D2,SEARCH(" ",D2)-1)),0)”*
- We calculated total minutes as *“=B1*60 + C1”*.

Unexpected characters appearing in the dataset were disregarded as recommended, considering them a temporary glitch.

3.3. Handling Missing Values

For missing values, we adopted different strategies based on the column's impact on the book's perception. Ratings with missing values were left as null to avoid unfairly influencing the book's reputation. Missing prices and times were filled with the median value of their respective columns, considering these don't significantly alter the perception of the book. All other missing values were left unchanged.

Analysis in Tableau

Analysing the release dates and total content duration, we observed a significant peak in 2021, coinciding with the COVID-19 pandemic when people had considerably more free time. This period likely saw an increased consumption of content, suggesting that Audible may have accelerated content publication to meet this heightened demand, a trend not observed in other years.

Then, we explored the top 5 authors and narrators based on the sum of stars in 2021. The rationale behind choosing the sum, rather than the average, stems from the observation that many authors and narrators boast an average rating of 5 stars, rendering the results somewhat inaccurate and not reflective of what we aimed to analyse. With the sum of stars, we can identify those with the highest levels of engagement on the site. Furthermore, we also showcased the sum of prices, enabling us to discern which authors and narrators generate the most revenue and merit further investment.

Additionally, it's noteworthy that some authors remained at the top in both the 2021 and 2022 graphs. This consistency could be attributed to their established fan base, which closely follows their work, suggesting that investing in more books from these authors could be profitable. Similarly, the emergence of new names in the 2022 graph indicates individuals earning recognition, meriting attention for potential investments. Of course, other factors must be considered before making investment decisions, but analysing engagement and revenue contribution offers a solid starting point.

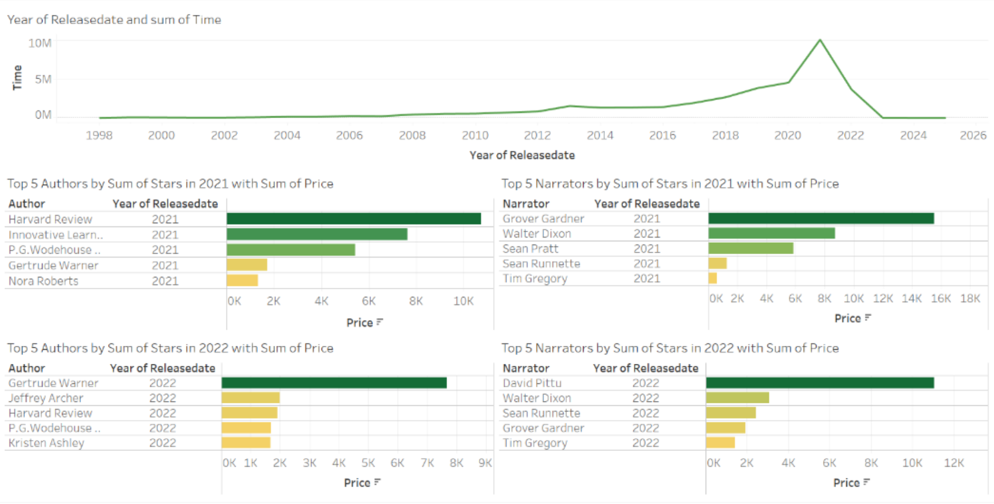

Fig 1. Audible Dashboard